基于改进注意力迁移的实时目标检测方法 *

张 弛[a, b]，刘宏哲[a, b†]
(北京联合大学 a. 北京市信息服务工程重点实验室; b. 机器人学院，北京 100101)

摘 要： 目前深度神经网络模型需要部署在资源受限的环境中，故需要设计高效紧凑的网络结构。针对设计紧凑的神经网络提出一种基于改进注意力迁移的模型压缩方法(KE)，主要使用一个宽残差教师网络(WRN)指导一个紧凑的学生网络(KENet)，将空间和通道的注意力迁移到学生网络来提升性能，并将该方法应用于实时目标检测。在 CIFAR 上的图像分类实验验证了经过改进注意力迁移的知识蒸馏方法能够提升紧凑模型的性能，在 VOC 上的目标检测实验验证了模型 KEDet 具有很好的精度(72.7mAP)和速度(86FPS)。实验结果充分说明基于改进注意力迁移的目标检测模型具有很好的准确性和实时性。

关键词： 神经网络；深度学习；目标检测；知识蒸馏；注意力迁移
中图分类号： TP311　　doi: 10.19734/j.issn.1001-3695.2020.02.0079

Real time object detection method based on improved attention transfer

Zhang Chi[a, b], Liu Hongzhe[a, b†]
(*a. Beijing Key Laboratory of Information Service Engineering*, *b. College of Robotics*, *Beijing Union University*, *Beijing* 100101, *China*)

Abstract: Recently, deep neural networks need to be deployed with low memory and computing resources, so it is necessary to design an efficient and compact network structure. This paper proposed a model compression method (KE) based on improved attention transfer for the design of compact neural networks, which mainly used a wide residual teacher network (WRN) to guide a compact student network (KENet) by extracting both spatial and channel-wise attention to improve the performance, and applied this method to real-time object detection. The image classification experiment on CIFAR verified that the knowledge distillation method with improved attention transfer can improve the performance of the compact model. The object detection experiment on VOC verified that the model KEDet has good accuracy (72.7mAP) and time performance (86FPS) . The experimental results show that the object detection model based on improved attention transfer has good accuracy and real-time performance.

Key words: neural network; deep learning; object detection; knowledge distillation; attention transfer

0 引言

目标检测是自动驾驶和辅助驾驶的重要环节，主要包括车辆检测、行人检测、交通标志检测、地面标志检测等任务。卷积神经网络(CNN)在目标检测任务中取得了显著的成绩，这很大程度上依赖于强大的计算能力和存储资源[1]。但在自动驾驶等领域往往资源受限，使得神经网络很难部署。因此，如何在保证性能的前提下有效地降低神经网络的计算和存储成本，是一个亟待解决的关键问题。

基于卷积神经网络的目标检测主要是以 R-CNN[2]系列为代表的基于区域提取(region proposal)的方法，也称为二阶段(two-stage)法。首先基于图片提出可能存在物体的候选区域(region of interest)，再通过候选区域来预测目标的类别的位置。这类方法检测精度较高，在大规模数据集上表现良好。然而这类方法往往计算量很大并且运行速度很慢，为了在保证准确率的同时减小计算量，提升运行效率和实时性，出现了以 YOLO[3]、SSD[4]为代表的单阶段(one-stage)目标检测框架。这类方法通过将图片划分成相同大小的网格，根据网格来预测目标的类别和位置。在此基础上的一系列新方法通过融合多尺度特征、联系上下文信息以及简化网络结构，进一步提升了目标检测的速度和精度[5]。

模型压缩是一类解决在资源受限条件下进行神经网络部署的通用方法[6]，可以对目标检测模型进行压缩来提升检测的实时性。对网络进行参数压缩的方法主要包括剪枝[7,8]、量化[9,10]和低秩分解[11,12]。除此之外，还可以通过更有效的卷积[13~15]设计更加紧凑的结构，或使用知识迁移[16~18](又称为知识蒸馏)，从一个大的“教师”模型中提取知识来帮助训练一个小的“学生”模型，这样可以提高“学生”模型的性能。注意力迁移(attention transfer)是一种改进的知识蒸馏(knowledge distillation)方法，通过将注意力机制引入知识蒸馏模型，并让学生网络和教师网络的注意力激活分布尽可能接近。

注意力迁移主要用于改进卷积神经网络，本文基于注意力迁移对轻量化卷积模型进行了改进，提取了空间和通道两个维度的知识，弥补了轻量化模型的不足，提出了名为知识增强的蒸馏方法。基于知识增强的方法，本文进一步提出了基于改进 SSD 的实时目标检测模型。通过在多个数据集上的实验，验证了基于改进注意力迁移的目标检测模型具有很好的准确性和实时性。

1 基于改进注意力迁移的知识蒸馏算法

1.1 轻量化卷积结构

现有的高性能目标检测模型往往使用更加轻量化的卷积

收稿日期： 2020-02-28；**修回日期：** 2020-04-23　**基金项目：** 国家自然科学基金资助项目(61871039、61802019、61906017)；北京市属高校高水平教师队伍建设支持计划资助项目(IDHT20170511)；北京联合大学领军人才项目(BPHR2019AZ01)；北京市教委项目(KM201911417001)；国家科技支撑计划资助项目(2015BAH55F03)；智能驾驶大数据协同创新中心(CYXC1902)；北京联合大学项目(WZ10201903)

作者简介： 张弛(1992-)，男，北京人，硕士研究生，主要研究方向为深度学习、计算机视觉；刘宏哲(1971-)，女(通信作者)，河北沧州人，教授，硕导，博士，主要研究方向为数字图像处理、人工智能(liuhongzhe@buu.edu.cn).

结构作为主干网络，下面来分析一下这些结构。在图 1(a)中，N 是输入通道数，K×K 是每个卷积核的大小，M 是输出通道数，总计算成本为 NK^2M。传统卷积的空间维数是卷积核的大小 K^2，而通道维数是输入和输出通道数 N×M。

减少参数的第一种方法是将每个卷积分成 G 组，如图 1(b)所示。与标准卷积 NK^2M 的计算成本相比，这种操作将计算量减少 1/G。在分组卷积之后使用 1×1 逐点卷积来提供一些跨通道的信息，它的计算成本为 N×M。总计算量为 NK^2M+NM。此方法已在 AlexNet [19]，Xnception [20] 和 ShuffleNet [14,15]中使用。

还有一种方法是使用窄而深的残差网络(如 ResNet，MobileNet 等)来代替宽而浅的神经网络(如 VGG16)，本质上是引入了瓶颈(bottleneck)结构。如图 1(c)所示，第一个 1×1 逐点卷积将输入通道维数 N 减小 B 倍，然后进行 K×K 卷积，最后，一个 1×1 逐点卷积恢复输出通道 M 的尺寸。总计算成本为 $NK^2M / B^2 + NK / B + MK / B$。

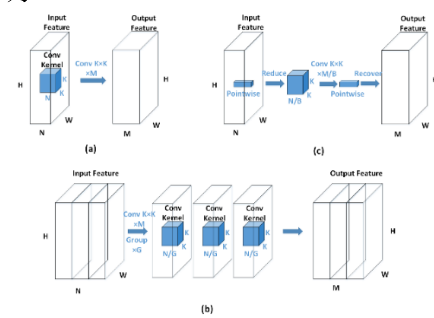


图 1　轻量化卷积结构

Fig. 1　Lightweight convolutional modules

基于以上两种方法本文提出了新的轻量化卷积结构：KENet。首先，将 1×1 分组卷积用于降维，然后执行传统的 3×3 卷积，最后，使用 1×1 逐点卷积恢复输出通道的尺寸。总计算成本为 $NK / GB + NK^2M / GB^2 + MK / B$。

使用分组卷积和瓶颈结构是减少参数的常用方法，但这也会带来信息的流失。如图 2(a)所示，从空间维度来看，窄而深的残差网络使用较小的卷积核，这会使感受野变小，从而丢失了一些空间上下文信息。如图 2(b)所示，从通道维度来看，使用分组卷积会将通道隔离，以使不同组之间的信息无法流通。典型的轻量化网络如 MobileNet 使用深度可分离卷积，它将每个通道都分为一组，过多的分组将大大降低运行速度。为了解决这些问题，可以在训练过程中，使用一个学习能力强的教师模型，通过添加额外的监督信号来进行知识迁移。

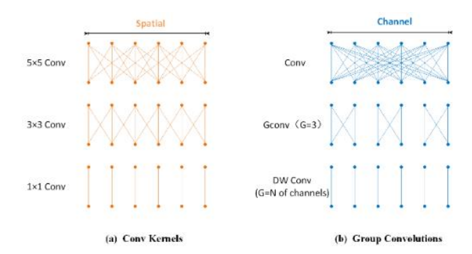


图 2　不同卷积核在空间和通道维度的投影

Fig. 2　Projection of different convolutional kernels in spatial and channel-wise dimensions

1.2　改进的注意力迁移算法

注意力迁移是一种改进的知识蒸馏方法。它采用了迁移学习的思想，将教师网络中间层的激活分布被视为源域，而学生网络的相应分布被视为目标域，通过最小化源域和目标域之间的距离，达到注意力迁移的目的。

注意力迁移通过激活图在中间层的一些注意点来测量学生与教师之间的距离，其损失函数定义如下：

$$\mathcal{L}_{AT} = \mathcal{L}(\sigma(s), x) + \frac{\beta}{2}\sum_{i=1}^{N}\left\| \frac{F(A_i^s)}{\|F(A_i^s)\|_2} - \frac{F(A_i^t)}{\|F(A_i^t)\|_2} \right\|_p \tag{1}$$

其中:s 和 t 分别表示学生和教师，$\sigma(\bullet)$ 是 softmax 函数，$\mathcal{L}(\sigma(\bullet), x)$ 表示标准交叉熵损失。i 表示从教师和学生中选取的激活层，$i = 1, 2, \cdots, N$。A_i^t 和 A_i^s 分别表示教师和学生的激活特征。$F(A_i)$ 表示注意力映射函数，它将三维的注意力激活张量映射为二维注意力激活图。使用 $F(A_i) = (1/N_i)\sum_j a_{ij}^2$ 作为映射函数，其中 a_{ij} 表示第 i 层中通道 j 的激活特征向量。β 是一个超参数，p 是范数类型，这里令 p=2。

为了解决轻量化结构的信息丢失问题，本文基于注意力迁移模型进行了改进，将空间和通道两个维度来提取知识，重新定义了注意力激活图，提出了称为知识增强的模型。对于激活特征图 $A_i \in R^{H \times W \times C}$，本文首先使用特征图通道间的关系来生成通道维度的知识。为了聚合空间信息，本文采用平均池化用于生成空间上下文描述符，然后将描述符输入一个全连接网络，以生成基于通道的知识 $K_c \in R^{1 \times 1 \times C}$。同时，本文利用特征图的空间关系来生成空间知识。首先沿通道轴使用平均池化操作以生成有效的特征描述符，然后输入一个卷积层产生空间知识 $K_s \in R^{H \times W \times 1}$。基于这两种知识，本文对注意力激活张量的知识进行了增强，并生成了增强后的激活特征 $E_i \in R^{H \times W \times C}$。完整的计算过程为

$$K_c(A_i) = \sigma(FC(Pool(A_i))) \tag{2}$$

$$K_s(A_i) = \sigma(Conv(Pool(A_i))) \tag{3}$$

$$E_i = K_s(A_i) \otimes K_c(A_i) \otimes A_i \tag{4}$$

其中，$\otimes$ 表示逐元素点乘。 最后，为教师和学生分别生成了知识增强激活图 $F(E_i)$。

改进后的注意力迁移模型如图 3 所示，由于该模型提取了空间和通道的知识来进行注意力迁移，弥补了轻量化卷积模型丢失的信息，故将其命名为知识增强模型。将知识增强模型部署在 KENet 等轻量化模型之中，仅仅增加了 3%左右的参数，额外的参数主要集中在产生空间知识的 FC 层之中，对于模型的整体参数而言可以忽略不计。

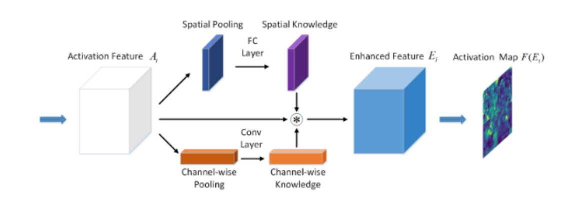


图 3　知识增强模型

Fig. 3　Knowledge enhance module

图 4 中可视化了在 ResNet50 的最后一个卷积层的激活图，对知识增强前后的效果进行了对比。红色部分表示强烈的激活并对最终结果作出了巨大贡献的部分。不难发现，在使用本文提出的知识增强模块后，网络更倾向于关注有用的区域。换句话说，通过增强空间知识和通道知识，本文所提出的方法使网络的注意力更加集中并提高了网络性能。

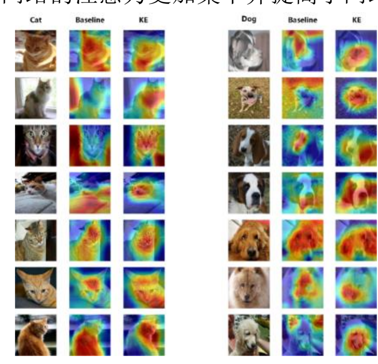


图 4　注意力激活图的可视化

Fig. 4　Visualization of the activation maps

2　基于改进 SSD 的实时目标检测模型

SSD(single shot multibox detector)是一个轻量级的单阶段目标检测算法，输入一张图片，经过 SSD 后直接生成分类和定位结果，实现了端到端的目标检测。SSD 的损失函数有如下形式：

$$L_{SSD}=\frac{1}{N}\sum_{i}L_{cls}+\alpha\frac{1}{N}\sum_{j}L_{loc} \tag{5}$$

其中 L_{cls} 是分类损失， L_{loc} 是定位损失，N 是正样本数，α 是平衡因子。本文基于知识蒸馏的思想对 SSD 进行了改进，整体框架结构如图 5 所示。

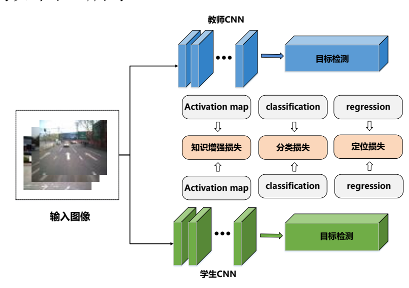


图 5　基于知识蒸馏的改进目标检测模型

Fig. 5　Improved object detection model based on knowledge distillation

其中将分类损失表示为

$$L_{cls}=\lambda L_{hard}(P_s,y_{cls})+(1-\lambda)L_{soft}(P_s,P_t) \tag{6}$$

其中 L_{hard} 是使用学生的真实标签 y_{cls} 预测的分类损失；$L_{soft}(P_s,P_t)=-\sum\omega_t P_t\log P_s$ 是使用教师和学生的软标签组成的蒸馏损失。定位损失可以表示为

$$L_{loc}=L_{smoothL1}(R_s,y_{loc})+\nu L_b(R_s,R_t,y_{loc}) \tag{7}$$

其中 $L_{smoothL1}$ 是使用真实标签 y_{loc} 和学生的位置预测的平滑 L1 损失，而 L_b 表示一个惩罚项，在学生的回归误差与教师的回归误差的差距超过一个界限 m 时进行惩罚，即

$$L_b=\begin{cases}\|R_s-y_{loc}\|_2^2 & \text{当}\|R_s-y_{loc}\|_2^2+m>\|R_t-y_{loc}\|_2^2\\ 0 & \text{其他}\end{cases} \tag{8}$$

此外，引入上文介绍的改进的注意力迁移(知识增强)算法，添加一个知识增强损失：

$$\mathcal{L}_{KE}=\sum_{i=1}^{N}\left\|\frac{F(E_i^s)}{\|F(E_i^s)\|_2}-\frac{F(E_i^t)}{\|F(E_i^t)\|_2}\right\|_p \tag{9}$$

其中 $F(E_i)$ 是知识增强激活图。最终改进的实时目标检测模型 KEDet 损失函数为

$$L_{KEDet}=\frac{1}{N}\sum_{i}L_{cls}+\alpha\frac{1}{N}\sum_{j}L_{loc}+\beta L_{KE} \tag{10}$$

在无人驾驶等领域，检测速度是确保实时性的关键，一般来说，模型越复杂检测的准确性越高，但检测速度就越慢。因此，需要在保证检测速度达到实时的基础上，尽量提升检测精度。SSD 这类单阶段检测算法可以在一定程度上达到实时，但却以牺牲精度为代价。本文提出的 KEDet 目标检测模型在 SSD 的基础上对主干网络进行了精简，使用 KENet 模型代替原来的 VGG 模型，提升了效率，并通过知识增强算法进一步提升了检测精度。

3　实验及结果分析

3.1　KENet 及知识增强算法的有效性

本文采用图像分类实验对知识增强的性能进行评估，实验中训练了三种类型的学生网络(Gconv、Bottleneck 和 KENet)，它们是从同一教师网络中提炼出来的。选用 Canadian Institute For Advanced Research (CIFAR)数据集进行图像分类实验，并使用官方指标(top-1 错误率)作为评估标准。

本文使用宽残差网络(WRN)作为实验的基本结构。它具有两个主要参数：深度 d 和宽度 k，其中深度 d 与卷积模块数 n 之间的关系为：d=6(n+4)，而宽度 k 决定了这些模块中滤波器的通道大小。宽残差网络的卷积部分由一个初始卷积层和三个主要卷积块组成。

实验使用带有标准残差模块的 WRN-40-2(宽残差网络的深度为 40，宽度乘数为 2)。每个标准模块由两个 3×3 卷积核组成。本文将使用以下的模块进行对比实验：

a) 分组卷积模块，命名为 Gconv(G)，G 是分组数，范围取{2,4,8,16}。

b) 具有 2 倍通道收缩的瓶颈模块，称为 Botteleneck(B)，其中 B＝2 是通道收缩倍数。

c) 典型的轻量级卷积神经网络 MobileNet，使用深度可分离卷积(Depthwise Separable Convolution, DSC)。

d) 本文设计了知识增强网络 KENet(G,B)将瓶颈结构与分组卷积结合在一起，使用 B = 2 且分组数 G 为{2,4,8,16}。

本文将知识增强(KE)方法与在没有知识蒸馏的情况下训练的模型进行了比较。使用了 4 个 Titan V GPU，minibatch 大小为 128，使用随机梯度下降对网络进行 200 个周期的训练，动量为 0.9，初始学习率为 0.1。每 60 个迭代将学习率降低 0.2 倍。超参数 β 设置为 1000。

表 1 给出了上述结构在 CIFAR10 数据集上进行图像分类的表现。首先比较不同卷积模块的计算成本，可以看出，通过分组卷积和瓶颈结构可以实现有效的参数压缩，KENet 将二者结合可以将参数压缩 10-20 倍。根据基于不同结构进行知识蒸馏得到的实验结果可以看出，使用本文提出的知识增强方法训练的学生模型的表现要明显优于直接训练的结果。当 KENet 作为学生并且使用知识增强作为蒸馏模型时，能够实现更有效的模型压缩，并且精度损失很小。由于 KENet 的参数远远少于教师和其他学生模型，这不可避免地导致准确性下降。由于知识增强模型提供了空间和通道信息，通过知识增强方法训练出的 KENet 的准确性有了显著的提升。

表 1　CIFAR10 上不同结构的分类误差

Tab. 1　Test error of different architectures on CIFAR10

block	architecture	complexity	baseline	KE
basic	Teacher: WRN-40-2	2238.6K	4.79	
Gconv(G)	GConv (2)	1357.4K	5.30	4.87
	GConv (4)	813.1K	5.50	5.00
	GConv (8)	541.0K	5.92	5.05
	GConv (16)	404.9K	6.65	5.13
Bottleneck(B)	Bottleneck (2)	430.5K	6.36	5.37
DWC	MobileNet	423.0K	8.61	6.64
KENet(G,B)	**KENet (2,2)**	**257.6K**	**7.14**	**5.57**
	KENet (4,2)	**170.2K**	**7.82**	**6.58**
	KENet (8,2)	**126.6K**	**8.40**	**6.83**
	KENet (16,2)	**104.7K**	**8.78**	**7.76**

3.2　实时目标检测模型的评估

为了验证本文提出的实时目标检测模型 KEDet 的有效性，将改进的 SSD 检测模型与原有 SSD 模型进行对比，并且选用二阶段检测模型 Faster-RCNN 也作为对比，评估了使用不同主干网络的检测性能。采用目标检测公共数据集 Pattern Analysis, Statistical Modelling and Computational Learning Visual Object Classes (PASCAL VOC)对模型的性能进行评估。使用平均检测准确率(mean average precision, mAP)来评估检测精度，并用每秒传输帧数(FPS)作为实时性的评价标准。

实验采用 VOC2007 和 VOC2012 作为训练数据集，输入图像大小为 300，使用具有 4 块 TITAN V GPU 的服务器训练 250 个 epoch，并在 VOC2007 测试集上进行性能评估，测试

设备是具有一块 GTX 1080 显卡的移动终端(笔记本电脑)。

表 2 给出了 Faster-RCNN、SSD 和 KEDet 模型的测试结果。Faster-RCNN 作为典型的二阶段检测模型，无论是基于 VGG16 还是 ResNet101 都具有较高的检测精度，但检测速度较慢，无法达到实时(30FPS 以上)。而 SSD(基于 VGG16)作为有代表性的单阶段检测模型，在保证检测精度的基础上达到了实时(45FPS)。但由于基于 VGG16 的 SSD 在实际部署的效果并不好，所以便产生了基于 MobileNet 的 SSD 目标检测模型，它牺牲了一定的检测精度(68.1mAP)，提升了检测速度(83FPS)，并且在实际部署中表现良好。然而，正如本文所述，MobileNet 模型使用了 Depthwise Separable Convolution，过多的分组导致了组间信息隔离，并且降低了运行速度。于是本文提出了 KENet 来代替 MobileNet 作为新的主干网络，降低了分组数，提升了一定的性能。

表 2　VOC2007 上不同结构的检测精度和速度对比

Tab. 2　Detection accuracy and speed comparison of different structures on VOC2007

detector	backbone	params	mAP	FPS
Faster-RCNN	VGG16	—	74.2	12
	ResNet101		76.4	7
SSD300	VGG16	26.5M	74.8	45
	MobileNet	5.8M	68.1	83
	KENet(2,2)	5.5M	68.5	86
KEDet	VGG16	30.2M	76.1	44
	MobileNet	6.0M	71.4	82
	KENet(2,2)	**5.7M**	**72.7**	**86**

与此同时，本文通过知识增强方法对检测模型进行了改进，提出了 KEDet 检测模型，通过改进的注意力迁移方法对原有模型进行知识蒸馏，从而弥补了轻量化模型丢失的信息。KENet 与 VGG16 相比，通过 bottleneck 结构使得网络层数更深，并且大大压缩了参数量。与 MobileNet 相比，用少量的分组代替了深度可分离卷积，提升了检测精度和速度。实验结果表明，KEDet 在提升检测精度(72.7mAP)的同时还保证了检测速度(86FPS)，具有很好的准确性和实时性。

4　结束语

本文提出了一种基于注意力迁移的知识蒸馏方法，并将其应用在 SSD 模型中进行了改进，提出了一个实时的目标检测模型 KEDet。首先通过分析轻量化卷积结构的特点，提出了一个结合分组卷积和瓶颈结构的轻量化卷积模型 KENet。然后提出了基于注意力迁移的知识蒸馏方法，对 KENet 模型进行了知识增强，并使用 CIFAR 数据集进行图像分类实验，验证了知识增强的有效性。基于此，提出了改进 SSD 的实时目标检测模型 KEDet，并在 VOC 数据集上进行了验证。实验结果表明，本文提出的 KEDet 模型具有较高的检测精度和检测速度，同时具备了准确性和实时性。未来的研究还可以结合剪枝算法以及神经架构搜索等，进一步探索更加高效的网络结构和压缩算法来提升检测效率。

参考文献：

[1] 张军阳，王慧丽，郭阳，等．深度学习相关研究综述 [J]．计算机应用研究，2018, 35(7): 1921-1928. (Zhang Junyang, Wang Huili, Guo Yang, *et al.* Review of deep learning [J]. Application Research of Computers, 2018, 35(7): 1921-1928.)

[2] Ren Shaoqing, He Kaiming, Girshick R, *et al.* Faster r-cnn: Towards Real-time Object Detection With Region Proposal Networks [C]// Advances in Neural Information Processing Systems. 2015: 91-99.

[3] Redmon J, Divvala S, Girshick R, *et al.* You only look once: Unified, real-time object detection [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[4] Liu Wei, Anguelov D, Erhan D, *et al.* SSD: Single shot multibox detector [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[5] 张琳娜，陈建强，陈晓玲，岑翼刚，阚世超．面向行车视频目标实时检测的轻量级 SSD 网络 [J]．计算机科学，2019, 46 (7): 233-237. (Zhang Linna, Chen Jianqiang, Chen Xiaoling, *et al.* Lightweight SSD network for real-time object detection in automotive videos [J]. Computer Science, 2019, 46 (7): 233-237.)

[6] 曹文龙，芮建武，李敏．神经网络模型压缩方法综述 [J]．计算机应用研究，2019(3): 649-656. (Cao Wenlong, Rui Jianwu, Li Min. Survey of neural network model compression methods [J]. Application Research of Computers, 2019(3): 649-656.)

[7] Yoon J, Hwang S J. Combined group and exclusive sparsity for deep neural networks [C]// Proc of the 34th International Conference on Machine Learning. 2017: 3958-3966.

[8] Liu Zhuang, Li Jianguo, Shen Zhiqiang, *et al.* Learning efficient convolutional networks through network slimming [C]// Proc of IEEE International Conference on Computer Vision. 2017: 2736-2744.

[9] Courbariaux M, Bengio Y, David J P. Binaryconnect: Training deep neural networks with binary weights during propagations [C]// Advances in Neural Information Processing Systems. 2015: 3123-3131.

[10] Hubara I, Courbariaux M, Soudry D, *et al.* Binarized neural networks [C]// Advances in Neural Information Processing Systems. 2016: 4107-4115.

[11] Wang Weiqi, Sun Yifan, Eriksson B, *et al.* Wide compression: Tensor ring nets [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9329-9338.

[12] Zhang Xiangyu, Zou Jianhua, He Kaiming, *et al.* Accelerating very deep convolutional networks for classification and detection [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2015, 38 (10): 1943-1955.

[13] Sandler M, Howard A, Zhu Menglong, *et al.* Mobilenetv2: Inverted residuals and linear bottlenecks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4510-4520.

[14] Zhang Xiangyu, Zhou Xinyu, Lin Mengxiao, *et al.* Shufflenet: An extremely efficient convolutional neural network for mobile devices [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6848-6856.

[15] Ma Ningning, Zhang Xiangyu, Zheng Haitao, *et al.* Shufflenet v2: Practical guidelines for efficient cnn architecture design [C]// Proc of European Conference on Computer Vision. 2018: 116-131.

[16] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network [EB/OL]. (2015-03-05) . https://arxiv. org/abs/1503. 02531.

[17] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer [EB/OL]. (2016-12-12) . https://arxiv. org/abs/1612. 03928.

[18] Yim J, Joo D, Bae J, *et al.* A gift from knowledge distillation: Fast optimization, network minimization and transfer learning [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4133-4141.

[19] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems. 2012: 1097-1105.

[20] Chollet F. Xception: Deep learning with depthwise separable convolutions [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1251-1258.